



Mathematisch-Naturwissenschaftliche Fakultät

Rafael R. C. Cuadrat | Danny Ionescu | Alberto M. R. Dávila |
Hans-Peter Grossart

Recovering Genomics Clusters of Secondary Metabolites from Lakes Using Genome-Resolved Metagenomics

Suggested citation referring to the original publication:
Frontiers in Microbiology 9 (2018) 251
DOI <https://doi.org/10.3389/fmicb.2018.00251>
ISSN (online) 1664-302X

Postprint archived at the Institutional Repository of the Potsdam University in:
Postprints der Universität Potsdam
Mathematisch-Naturwissenschaftliche Reihe ; 924
ISSN 1866-8372
<https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-445656>
DOI <https://doi.org/10.25932/publishup-44565>



Recovering Genomics Clusters of Secondary Metabolites from Lakes Using Genome-Resolved Metagenomics

Rafael R. C. Cuadrat^{1,2,3*}, Danny Ionescu², Alberto M. R. Dávila⁴ and Hans-Peter Grossart^{2,5}

¹ Bioinformatics Core Facility, Max Plank Institute for Biology of Ageing, Köln, Germany, ² Experimental Limnology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Stechlin, Germany, ³ Berlin Center for Genomics in Biodiversity Research, Berlin, Germany, ⁴ Computational and Systems Biology Laboratory, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil, ⁵ Institute of Biochemistry and Biology, Potsdam University, Potsdam, Germany

OPEN ACCESS

Edited by:

Diana Elizabeth Marco,
National Scientific Council (CONICET),
Argentina

Reviewed by:

Hans Uwe Dahms,
Kaohsiung Medical University, Taiwan
Steven Singer,
Lawrence Berkeley National
Laboratory (LBNL), United States

*Correspondence:

Rafael R. C. Cuadrat
rafaelcuadrat@gmail.com

Specialty section:

This article was submitted to
Aquatic Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 24 October 2017

Accepted: 31 January 2018

Published: 20 February 2018

Citation:

Cuadrat RRC, Ionescu D, Dávila AMR
and Grossart H-P (2018) Recovering
Genomics Clusters of Secondary
Metabolites from Lakes Using
Genome-Resolved Metagenomics.
Front. Microbiol. 9:251.
doi: 10.3389/fmicb.2018.00251

Metagenomic approaches became increasingly popular in the past decades due to decreasing costs of DNA sequencing and bioinformatics development. So far, however, the recovery of long genes coding for secondary metabolites still represents a big challenge. Often, the quality of metagenome assemblies is poor, especially in environments with a high microbial diversity where sequence coverage is low and complexity of natural communities high. Recently, new and improved algorithms for binning environmental reads and contigs have been developed to overcome such limitations. Some of these algorithms use a similarity detection approach to classify the obtained reads into taxonomical units and to assemble draft genomes. This approach, however, is quite limited since it can classify exclusively sequences similar to those available (and well classified) in the databases. In this work, we used draft genomes from Lake Stechlin, north-eastern Germany, recovered by MetaBat, an efficient binning tool that integrates empirical probabilistic distances of genome abundance, and tetranucleotide frequency for accurate metagenome binning. These genomes were screened for secondary metabolism genes, such as polyketide synthases (PKS) and non-ribosomal peptide synthases (NRPS), using the Anti-SMASH and NAPDOS workflows. With this approach we were able to identify 243 secondary metabolite clusters from 121 genomes recovered from our lake samples. A total of 18 NRPS, 19 PKS, and 3 hybrid PKS/NRPS clusters were found. In addition, it was possible to predict the partial structure of several secondary metabolite clusters allowing for taxonomical classifications and phylogenetic inferences. Our approach revealed a high potential to recover and study secondary metabolites genes from any aquatic ecosystem.

Keywords: metagenomics 2.0, PKS, NRPS, freshwater, environmental genomics

INTRODUCTION

Metagenomics, also known as environmental genomics, describes the study of a microbial community without the need of *a priori* cultivation in the laboratory. It has the potential to explore uncultivable microorganisms by accessing and sequencing their nucleic acid (Rodríguez-Valera, 2004). In recent years, due to decreasing costs of DNA sequencing—metagenomic databases

(Vincent et al., 2016) (e.g., MG-RAST) have rapidly grown and archive billions of short read sequences (Meyer et al., 2008). Many metagenomic tools and pipelines were proposed to better analyse these enormous datasets (Huson et al., 2007). Additionally, these tools allow to (i) infer ecological patterns, alpha-, and beta-diversity and richness (Caporaso et al., 2010); (ii) assemble environmental contigs from the reads (Li et al., 2015), and more recently, (iii) recover draft genomes from metagenomic bins (Strous et al., 2012; Wu et al., 2014; Kang et al., 2015). By recovering a high number of draft genomes from these so far uncultivable organisms, it is now possible to screen for new genes and clusters, unlocking a previously underestimated metabolic potential such as secondary metabolite gene clusters by using a metagenomic approach called Metagenomics 2.0 (McMahon, 2015).

Lake Stechlin is an oligo-mesotrophic ecosystem located in the Brandenburg-Mecklenburg lake district, northeast Germany, in a glacial region with numerous dead-ice lakes greatly varying in water quality according to the European Water Framework Directive (WFD) (Dadheech et al., 2014). The lake is monitored since 1959 for several physical, chemical and biological variables, ranging from phyto- and zooplankton species composition to cell number and biomass calculations (Jost Casper, 1985), but only since the late 20th, early 21st century mass developments of cyanobacteria were found (Salmaso and Padišák, 2007).

Many species of Cyanobacteria can form blooms, and it has been estimated that 25–75% of cyanobacterial blooms are toxic, specially in warm water (Bláhová et al., 2007, 2008) and the frequency of these blooms has also risen in Lake Stechlin in recent decades (Jost Casper, 1985) whereby its toxicity seems to increase in the lake (Dadheech et al., 2014).

Polyketide synthases (PKS) and non-ribosomal peptide synthases (NRPS) are two families of modular mega-synthases, both are very important for the biotechnological and pharmaceutical industry due to their broad spectrum of products, spanning from antibiotics and antitumor drugs to food pigments and also harmful toxins, like Anatoxin-a, and Microcystins (Dadheech et al., 2014).

Both families of mega-synthases act in an analogous way, producing polyketides (using acyl-coA monomers) and peptides (using aminoacyl monomers), and they are broadly distributed in many taxonomical groups, ranging from bacteria (alphaproteobacteria, cyanobacteria, actinobacteria) to fungi (Gokhale et al., 2007; Koglin and Walsh, 2009).

PKS enzymes can be classified in types (I, II, and III), whereby type I can be further classified into modular or iterative classes. The iterative PKS use the same domain many times, iteratively, to synthesize the polyketide. The modular PKS are large multi-domain enzymes in which each domain is used only once in the synthesis process (Cane et al., 1998; Lal et al., 2000). The production of the polyketide follows the co-linearity rule, each module being responsible for the addition of one monomer to the growing chain (Minowa et al., 2007).

Type I PKS are characterized by multiple domains in the same open reading frame (ORF) while in type II each domain is encoded in a separate ORF, acting interactively (Sun et al., 2012). Type III is also known as Chalcone synthase and has different evolutionary origin from type I and II (Austin and Noel, 2003).

Type III PKSs are self-contained enzymes that form homodimers. Their single active site in each monomer catalyzes the priming, extension, and cyclization reactions iteratively to form polyketide products (Austin and Noel, 2003). Hybrid PKS/NRPS and NRPS/PKS are also modular enzymes, encoding lipopeptides (hybrid between polyketides and peptides) and occur in bacterial as well as fungal genomes (Fisch, 2013; Masschelein et al., 2013; Mizuno et al., 2013).

PKS and NRPS are very well explored in genomes from cultivable organisms, mainly *Actinomycetes* (Komaki et al., 2014) and *Cyanobacteria* (Micallef et al., 2015). Recently, by using a metagenomic approach, studies have demonstrated the presence of these metabolite-genes in aquatic environments, as for example, Brazilian coastal waters (in free living and particle-associated bacteria) and from the microbiomes of Australian marine sponges (Woodhouse et al., 2013). However, there are few metagenomic studies whose scope is to find these gene families in freshwater environments where most studies are based on isolation approaches (Silva-Stenico et al., 2011; Zothanpuia et al., 2016).

In addition, due to the rather large size of genes involved in these pathways, yet, it is not possible to recover the full genes by using traditional read-based metagenomics or the single sample assembly approach. Most of the studies aim to solely find specific domains, like Keto-synthase (KS) in PKS and Condensation domain (C) in NRPS, due to the high conservation of these domains (Selvin et al., 2016).

Thus, it is important to develop new methods to recover full sequences of these families, especially because of their modular nature, i.e., the final compound can only be inferred with information about all the modules and domains based on the co-linearity role of polyketides and non-ribosomal peptides (Minowa et al., 2007). By detecting the full sequences, the approach can be used, for example, for environmental long-term monitoring of the toxicity potential of harmful algal blooms (HABs) (Meriluoto et al., 2017).

We used a metagenomics 2.0 approach to overcome these limitations and improve the screening for secondary metabolism genes and clusters while evaluating the potential of microbial communities for future research on potential drugs and toxin production. This study aims to (i) generate draft genomes from Lake Stechlin; (ii) to screen these genomes for new complete multi-modular enzymes from PKS and NRPS families, exploring their diversity and phylogeny.

MATERIALS AND METHODS

Sampling and Sequencing

A total of 26 metagenomic samples from Lake Stechlin, north-eastern Germany were used.

Water was collected as metagenomic samples on several occasions (April, June 2013, July 2014, August 2015) in sterile 2 L Schott bottles from Lake Stechlin (53°9′5.59N, 13°1′34.22E). All samples, except those from August 2015, were filtered through 5 μm and subsequently 0.2 μm pore-size filters. The samples collected in August 2015 were not size-fractionated and directly filtered on a 0.2 μm pore size filter. Genomic DNA was extracted

using a phenol/chloroform protocol as described in Ionescu et al. (2012) and was sent for sequencing.

Sequencing was conducted at MrDNA (Shallowater, Texas) on an Illumina HiSeq 2500, using the V3 chemistry, following, fragmentation, adaptor ligation and amplification of 50 ng genomic DNA from each sample, using the Nextera DNA Sample Preparation Kit.

Table S1 shows the general information about the 26 samples used in this study.

Environmental Draft Genomes

Briefly, all samples were pre-processed by Nesoni (<https://github.com/Victorian-Bioinformatics-Consortium/nesoni>) to remove low quality sequences and to trim adaptors, and afterwards assembled together using MegaHIT (default parameters) (Li et al., 2015). The reads from each sample were mapped back to these assembled contigs using BMAP (<https://sourceforge.net/projects/bbmap/>) and then all data was binned using MetaBAT (Kang et al., 2015) to generate the draft genomes. The completeness and taxonomical classification were checked using CheckM (Parks et al., 2015).

Screening Secondary Metabolism Genes and Phylogenetic Analysis of NRPS and PKS Domains

DNA fasta files of the generated bins (288) were submitted to a locally installed version of Anti-SMASH (`-clusterblast -smcogs -limit 1500`) (Weber et al., 2015). Using in-house ruby scripts, the domains from PKS and NRPS were parsed. The PKS KS domains and NRPS C domains were submitted to NAPDOS for classification (Ziemert et al., 2012). In addition, all the KS and C domains (trimmed by NAPDOS) were submitted to BLASTP against RefSeq database (O'Leary et al., 2016), using the default parameters. The 3 best hits of each domain were extracted and added to the original multi-fasta file with the environmental domains. The full set of KS and C domains (from bins and references obtained by the blast on RefSeq database) was submitted for NAPDOS for the phylogenetic analysis. The resulting alignments and trees were exported, then trees were manually checked and annotated.

Relative Abundance of Bins in Each Sample

The reads from each sample were mapped (using BMAP) against each bin fasta file and an in-house ruby parser script was used to calculate the relative abundance of each bin in each sample, normalizing the read counts by the number of reads of each sample. The table with the results was loaded into STAMP (Parks et al., 2014) in order to analyse the significant differences of bin abundance over the samples.

RESULTS

All the sequences generated for this study were submitted to ENA under accession numbers: PRJEB22274 and PRJEB7963.

Environmental Draft Genomes Obtained (Bins)

Metagenomic binning resulted in 288 draft environmental genomes (called bins in this study). Of these, 45 had a predicted completion level higher than 75% according to CheckM.

Table S2 shows the general information about each bin, including completeness, genome size, number of open read frames (ORFs) and taxonomical classifications (from CheckM).

Screening Secondary Metabolism Genes and Phylogenetic Analysis

By using Anti-SMASH, at least one secondary metabolite gene cluster was found in 121 of the bins, totaling 243 clusters and 2200 ORFs. From these 243 clusters, 125 (51.4%) were classified in the Terpene and 35 (14.40%) in the bacteriocin pathway. In addition, a total of 18 NRPS, 6 type I PKS and 3 hybrid PKS/NRPS clusters were found in 15 different bins (**Figure 1A**). The latest 3 obtained pathway clusters are the main focus of our study.

Figure 1B shows the taxonomical classification at phylum level for the bins showing NRPS, type I PKS and hybrid clusters. **Table S3** shows the distribution of all clusters in all bins.

A total of 43 condensation (C) domains were obtained from NRPS clusters. All these sequences were submitted to NAPDOS analysis. **Figure 2A** shows the classification of C domains into classes. Most of the sequences were classified as LCL domains (58%). This kind of domain catalyzes the formation of a peptide bond between two L-amino acids.

The screening for type I PKS resulted in 9 KS domain sequences. Most of them are classified as modular type I PKS (56%). All of them were submitted to NAPDOS and classified into 4 different classes (**Figure 2B**).

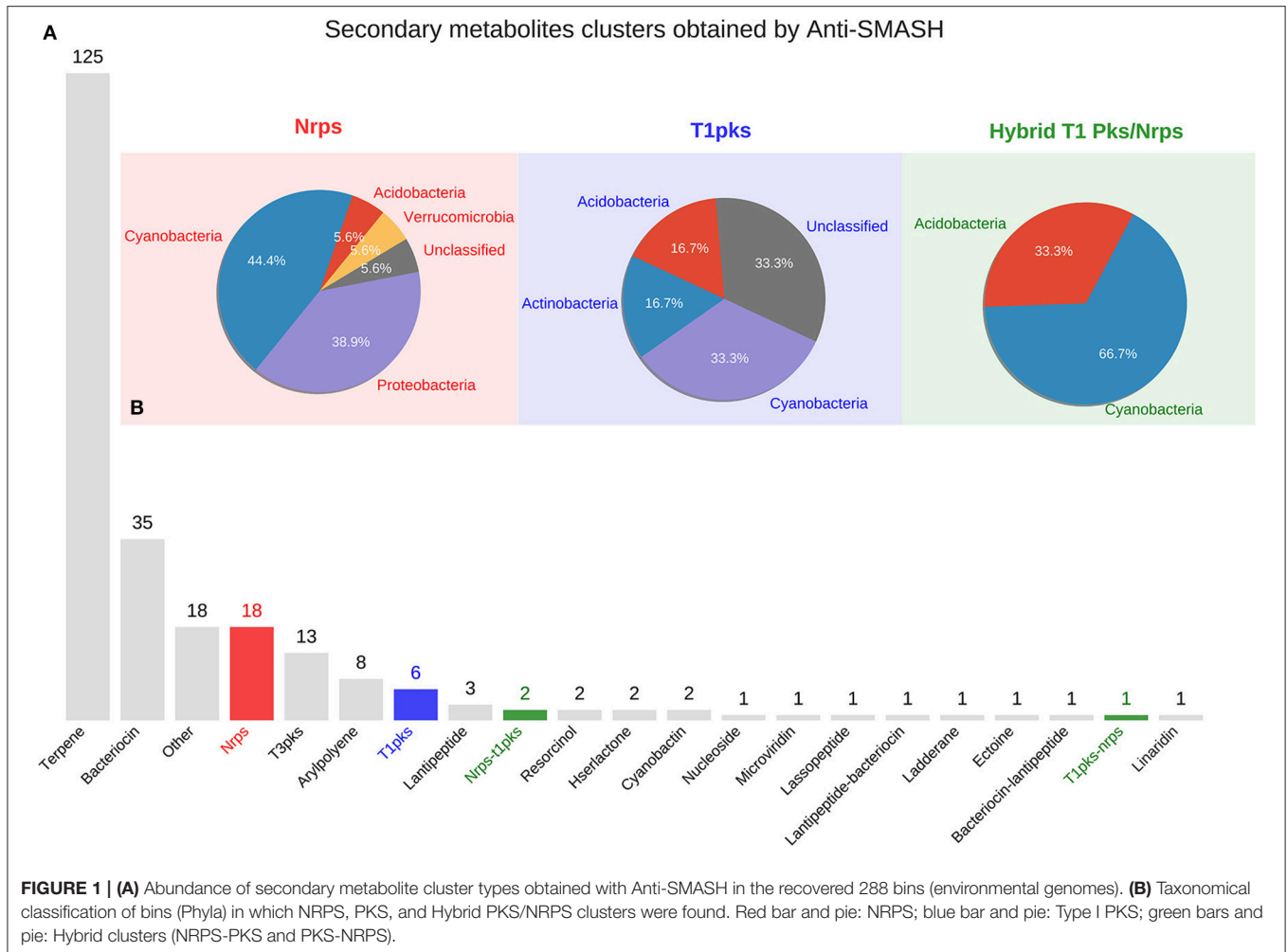
All the KS and C domains were also submitted to similarity analysis by using BLASTP against RefSeq database (**Tables S4, S5**) and the best 3 hits of each sequence were extracted and used for phylogenetic analyses with NAPDOS. The trees for C and KS domains are shown in **Figures 3 and 4**, respectively.

Relative Abundance of Bins on Each Sample

The relative abundance of all the bins in each sample was estimated by mapping the reads from each sample against the assembled bins. **Table S6** shows the normalized bin abundance for every sample.

Due to the differences between the filtration methods, we decided to classify the samples in 3 groups: particle associated samples (PA)-filtered on 5 μm membranes (and also samples from aggregates), free-living samples (FL)-pre-filtered through 5.0 μm membranes and subsequently filtered on 0.22 μm membranes, and non-size fractionated samples (NSF)-filtered direct on 0.22 μm membranes (without previous filtering) retaining the whole bacterial community. In total, we obtained 7 samples in the PA group, 5 in the FL group, and 14 in the NSF group.

The table with the relative abundance of the bins in all samples was loaded on STAMP and an ANOVA test was conducted followed Games-Howell *POST-HOC* test and



Benjamini-Hochberg FDR correction. **Table S7** shows 158 bins for which the difference in relative abundance was statistically significant ($p < 0.05$) between the 3 groups (FL, PA, and NSF).

From the 15 bins containing NRPS and/or type I PKS clusters, only 4 showed significant difference between the 3 groups. Bins 1 and 2 are more abundant in PA samples and bins 193 and 235 are more abundant in the FL samples.

Exploring NRPS, Type I PKS, and Hybrid Clusters from Draft Genome Bins

We highlight 3 bins (with less than 35% contamination and more than 70% completeness) out of the 15 obtained type I PKS and/or NRPS and explore their clusters.

In bin 34 (*Pseudomonas*, 98.28% completeness) it was possible to retrieve 7 clusters, including 3 NRPS clusters (**Figure 5**) and 2 bacteriocin clusters.

In cluster 2 (ctg181), multiple domains of NRPS (with the 3 minimal modules) and regulatory genes were identified, e.g., smCOG: SMCOG1057 (TetR family transcriptional regulator) (**Figure 5**, in green arrows).

All the clusters show a high similarity with *Pseudomonas* proteins. Cluster 2 has a similarity of 92% with *Pseudomonas synxantha* bg33r, conserving also the gene synteny.

The C domain sequences were submitted to NAPDOS analysis and 2 were classified as belonging to the syringomycin pathway and the LCL class, and one was classified as belonging to the microcystin pathway or and the DCL class (link an L-amino acid to a growing peptide ending with a D-amino acid).

In cluster 3 (ctg415) (**Figure 5**), in addition to the NRPS domains, the following transporter related genes were found: smCOG: SMCOG1288 (ABC transporter related protein) and SMCOG1051 (TonB-dependent siderophore receptor) (blue narrow arrows). Nevertheless, this cluster is not complete and just one C domain was found (LCL class), which was also classified to the syringomycin pathway.

Cluster 6 (ctg857–**Figure 5**) shows many NRPS domains, regulatory factors and transporters genes, including drug resistance genes, e.g., SMCOG1005 (drug resistance transporter, EmrB/QacA), SMCOG1044 (ABC transporter, permease protein) and SMCOG1051 (TonB-dependent siderophore receptor) (blue arrows). Two C domains from this cluster were classified as belonging to the heterocyclization class. This class catalyzes

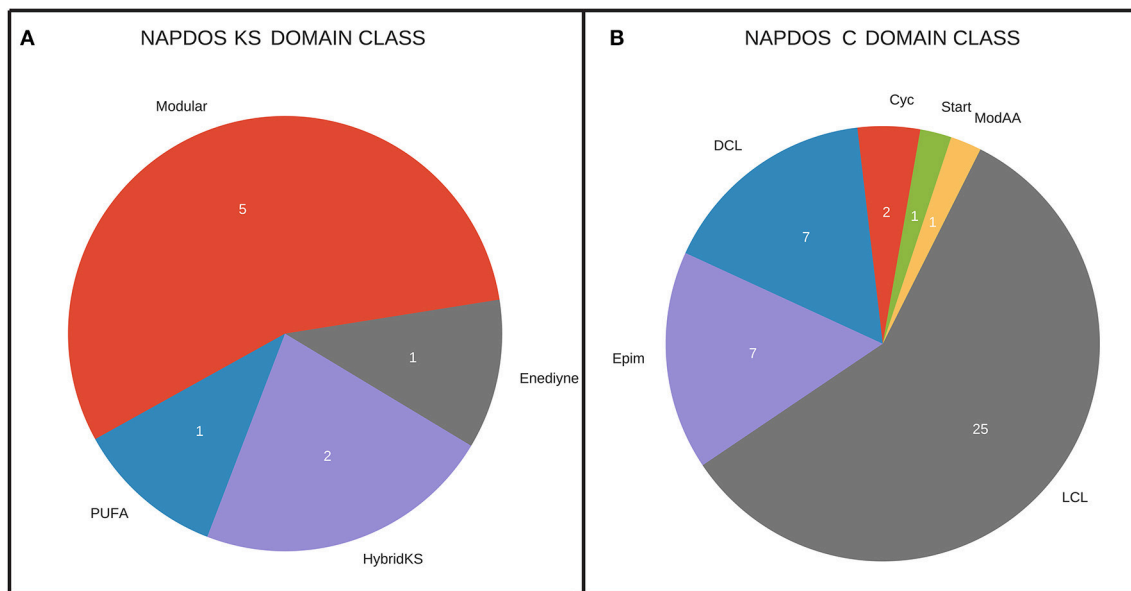


FIGURE 2 | (A) NAPDOS classification of the NRPS KS domain. Modular: possess a multidomain architecture consisting of multiple sets of modules; hybridKS: are biosynthetic assembly lines that include both PKS and NRPS components; PUFA: Polyunsaturated fatty acids (PUFAs) are long chain fatty acids containing more than one double bond, including omega-3-and omega-6- fatty acids; Eneidyne: a family of biologically active natural products. The enediyne core consists of two acetylenic groups conjugated to a double bond or an incipient double bond within a nine- or ten-membered ring. **(B)** NAPDOS classification of NRPS C domain. Cyc, cyclization domains catalyze both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues; DCL, link an L-amino acid to a growing peptide ending with a D-amino acid; Epim, epimerization domains change the chirality of the last amino acid in the chain from L- to D-amino acid; LCL, catalyze formation of a peptide bond between two L-amino acids; modAA, appear to be involved in the modification of the incorporated amino acid; Start, first module of a Non-ribosomal peptide synthase (NRPS).

both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues (Di Lorenzo et al., 2008). Both domains were classified in the pyochelin pathway by NAPDOS. The phylogenetic tree (Figure 3) confirms both the functional and taxonomical classification (confidence value 100).

In bin 193 (Mycobacterium, 73.37 % completeness) a type I PKS cluster was identified (ctg514) (Figure 6). Five PKS domains were retrieved, including the minimal core from one of the ORFs on this contig. The KS domain BLASTP result shows 82% (and 99% coverage) similarity with *Mycobacterium kansasii*. The NAPDOS analysis from the KS domain suggests that it could be a modular (epothilone pathway) or iterative type I PKS similar to the calicheamicin pathway. However, by using the phylogenetic analysis it was clustered as the iterative clade (confidence value 98.4) together with the *M. kansasii* sequence (confidence value 100) (Figure 4).

Two further clusters were recovered: one type III PKS and one unclassified one. All clusters show similarity with the *Mycobacterium* clusters.

Bin 131 (unclassified bacteria by CheckM) has 84.09% of completeness reported by CheckM. In this bin it was possible to find one cluster and 3 domains of type I PKS (KS, AT, and KR) (Figure 6). The KS domain was classified by NAPDOS as belonging to the maduropeptin and neocarzinostatin pathways. Using clusterblast inside Anti-Smash it was not possible to

find any similar cluster, but using BLASTP it was possible to find similarity with the cyanobacteria *Microcystis aeruginosa* (64% identity and 99% coverage on BLASTP search). In the phylogenetic tree, it was clustered within the enediynes clade (Figure 4) and also with *M. aeruginosa* (confidence value 99).

In addition, there were 12 more bins with NRPS or type I PKS clusters, but with less than 70% of completeness or more than 35% contamination. The bins 1 (69.54% completeness) and 2 (16.52%) were classified as the genus *Anabaena*, showing NRPS and hybrid NRPS-type 1 PKS, respectively. Bins 6 (39.66% completeness), 7 and 8 were classified as the genus *Planktothrix* and show a high diversity of secondary metabolites: 3 NRPS clusters, one type I PKS and 2 NRPS-PKS hybrid (bin 6), 2 NRPS (bin 7) and 2 NRPS and one type I PKS (bin 8). The bin 8 also shows a microviridin cluster.

Bins 73 and 217 are classified as *Acidobacteria* showing PKS and NRPS, respectively. Bin 235 (*Burkholderiaceae* family) shows 3 NRPS clusters and bin 13 (*Comamonadaceae* family, also from *Burkholderiales* order) shows one NRPS cluster.

Bin 78 is classified as *Verrucomicrobiaceae* shows 1 NRPS cluster. Additionally, there is an unclassified *Archaea* (NRPS cluster) and one bin without any classification (bin 136, type I PKS).

The Anti-Smash results for all bins are available in the Supplemental Information (Table S1).

DISCUSSION

The field of metagenomics has generated a vast amount of data in the last decades (National Research Council US., 2007). Most of the data is poorly annotated and may show little quality control when loaded into the public databases, hence awaiting a more in-depth analysis (Gilbert et al., 2011). There are many open challenges in this field, e.g., (i) the lack of representative genomic databases from uncultivable organisms to be used in a similarity-based annotation procedure; (ii) high confidence assembly of short reads from species-rich samples; (iii) obtaining high enough coverage for every organism in the sample, including those with a low abundance, etc. (Teeling and Glöckner, 2012). Recently, some new algorithms have been proposed to overcome these limitations and to obtain partial or near complete genomes from environmental samples, e.g., MetaBat (Kang et al., 2015) and MetaWatt (Strous et al., 2012). Most of them require many samples and high coverage sequencing per sample as an input. Recently, studies have been done to recover genomes even from rare bacteria (Albertsen et al., 2013). The term Metagenomics 2.0 was introduced to describe this new generation of metagenomic analysis by Katherine McMahon (McMahon, 2015) and most of the studies using this approach have been conducted to reveal ecological interactions and networks (Sangwan et al., 2016; Vanwonterghem et al., 2016).

In this study, we recovered 288 environmental draft genomes using 26 samples from Lake Stechlin, a temperate oligomesotrophic lake. One of the advantages of this approach is to enable the recovery of large genomic clusters, especially the megasynthases clusters of the secondary metabolism, e.g., involved in biosynthesis of antibiotics, including its regulatory and transporter genes. Here, we have used the Anti-SMASH and NAPDOS pipelines to identify, annotate, classify and to carry out the phylogenetic analysis of a total of 243 clusters of known secondary metabolites. To our knowledge, this is the first study using the metagenomics 2.0 approach to recover megasynthases clusters. A number of previous studies had been conducted using a traditional PCR based screening (Amos et al., 2015) and shotgun metagenomic approach (Foerstner et al., 2008; Cuadrat et al., 2015) exploring the abundance and diversity of individual genes and domains, but these studies are missing the genomic context. By obtaining the entire genomic context it is possible, in future studies, to clone and to do heterologous expression for all the genes, including promoters and transporters.

The phytoplankton from Lake Stechlin has been intensively investigated since 1959, allowing researchers to follow long-term trends, for example, mass developments of cyanobacteria, found only since the late 20th, early 21st century (Salmaso and Padišák, 2007), coinciding with an increase in surface temperature (by 0.37°C per decade to an annual mean of 11.3 ± 0.5°C) (Kirillin, 2013).

The increase in water temperature has been paralleled by the occurrence of HABs, favoring potentially toxic cyanobacteria (Paerl and Huisman, 2008) such as *Microcystis*, which has been found in the Stechlin Lake since 2011 (Dadheech et al., 2014).

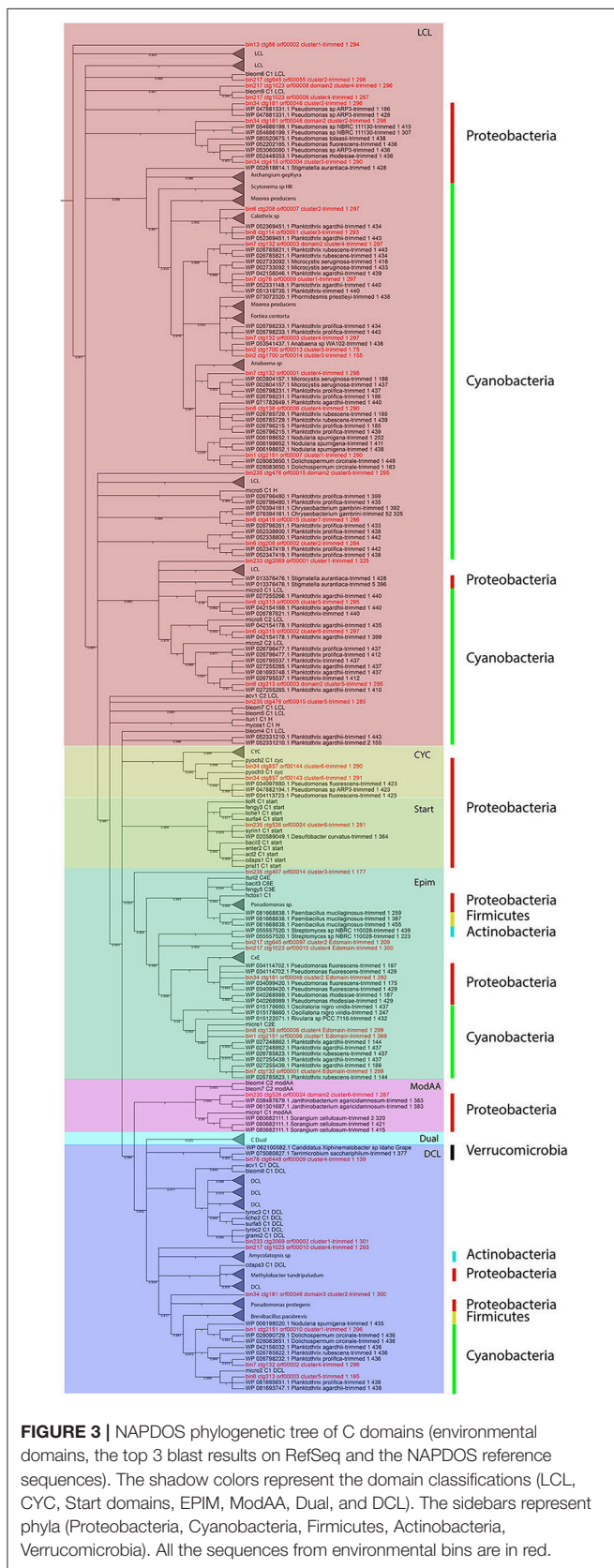


FIGURE 3 | NAPDOS phylogenetic tree of C domains (environmental domains, the top 3 blast results on RefSeq and the NAPDOS reference sequences). The shadow colors represent the domain classifications (LCL, CYC, Start domains, EPIM, ModAA, Dual, and DCL). The sidebars represent phyla (Proteobacteria, Cyanobacteria, Firmicutes, Actinobacteria, Verrucomicrobia). All the sequences from environmental bins are in red.

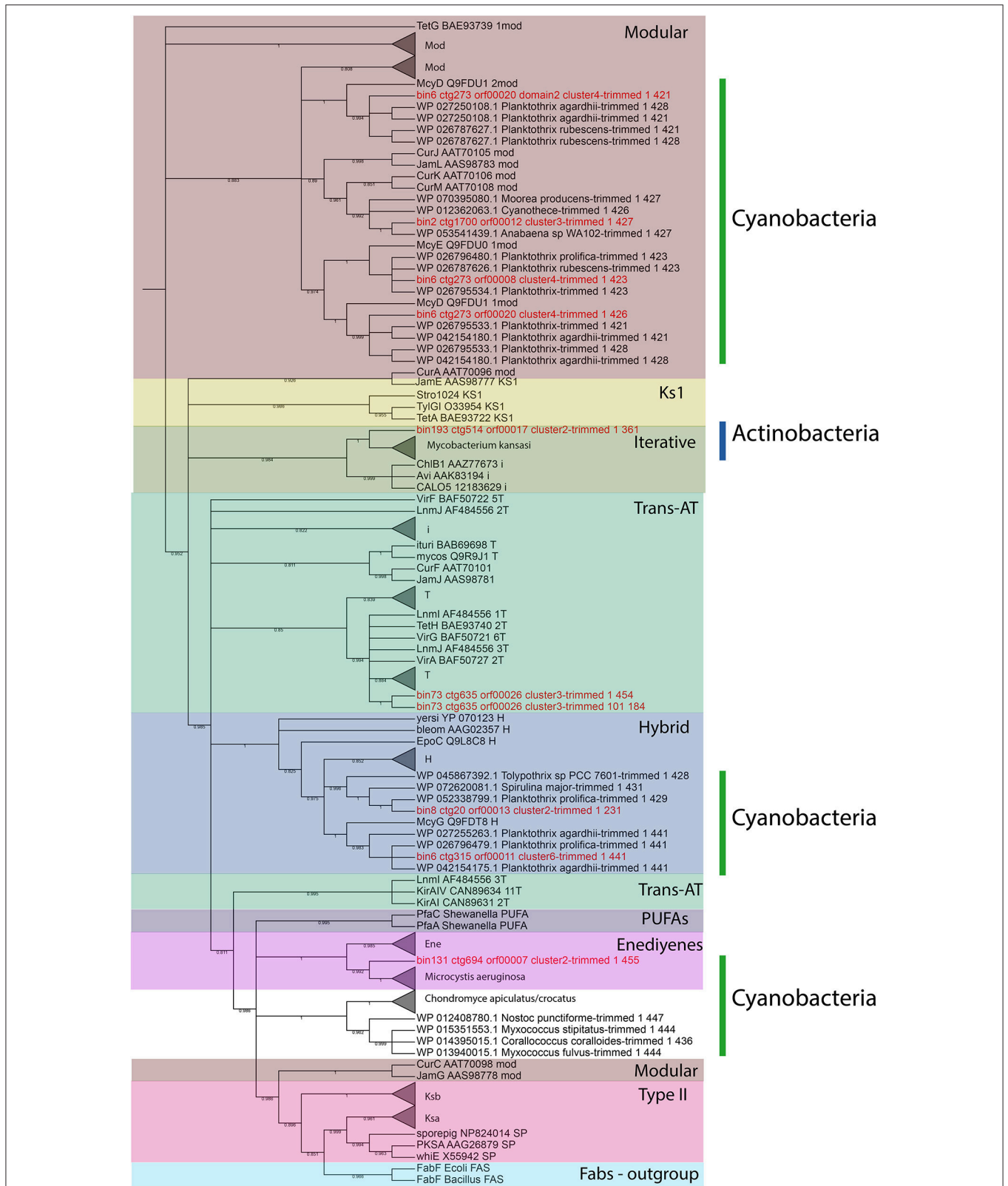


FIGURE 4 | NAPDOS tree of KS domains (environmental domains, the top 3 blast results on RefSeq and the NAPDOS reference sequences). The shadow colors represent the domain classifications (Modular, KS1, Iterative, Trans-AT, Hybrid, PUFA, Enediynes, Type II, and Fabs, Fatty acid synthase). The sidebars represent phyla (Cyanobacteria and Actinobacteria). All the sequences from environmental bins are in red.

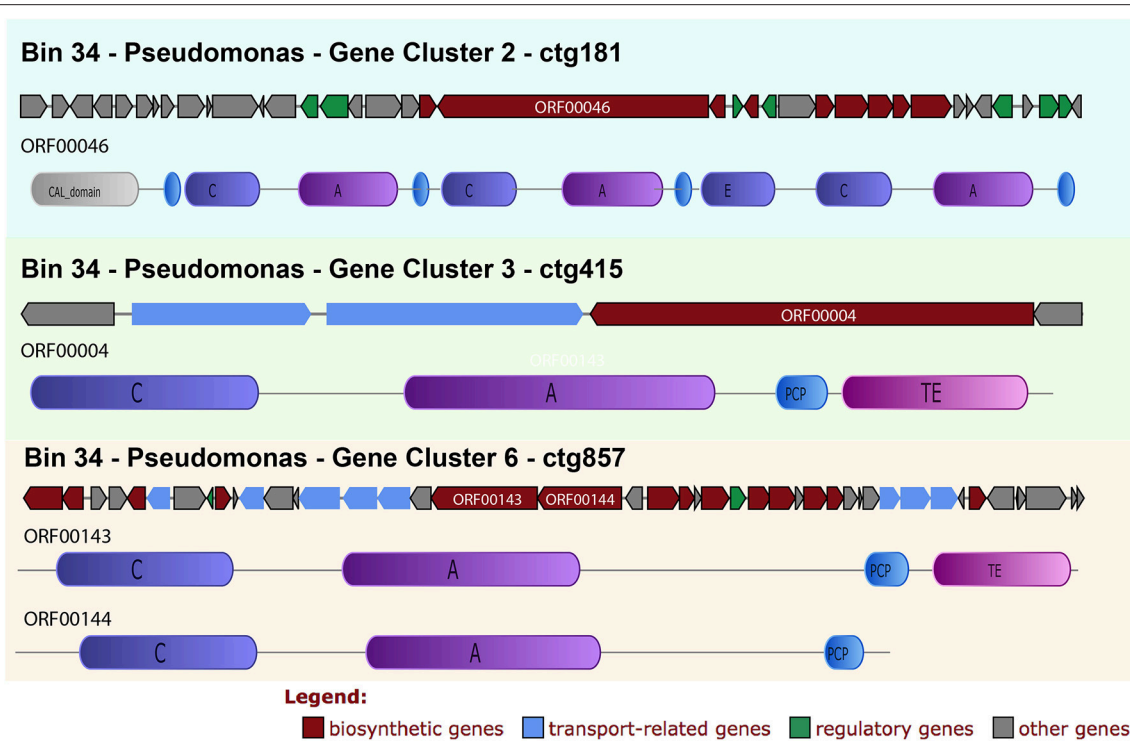


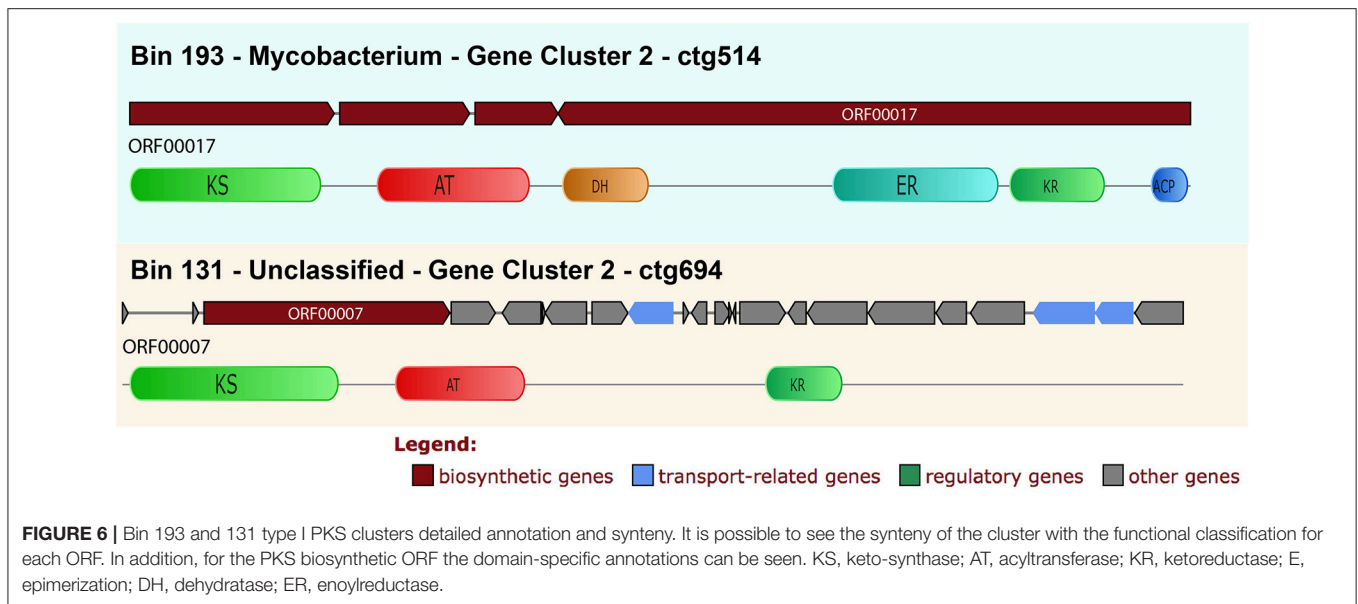
FIGURE 5 | Bin 34, NRPS clusters detailed annotation and synteny. The synteny of the clusters with a functional classification for each ORF is given. In addition, for the NRPS biosynthetic ORFs the domain annotations are given. CAL, Co-enzyme A ligase domain; C, condensation; A, adenylation; E, epimerization; TE, Termination; KR, Ketoreductase domain; ECH, Enoyl-CoA hydratase.

In the same study, for the first time, the presence of microcystin genes has been detected in the lake. Microcystins represent a family of genes coding one of the most common hepatotoxins that are produced by a wide range of cyanobacteria (Rantala et al., 2008; Ballot et al., 2010b). The genes encoding for these toxins belong to the secondary metabolism pathways, mostly the PKS, and NRPS families.

Screening the bins for secondary metabolite clusters, we see that the most abundant cluster belongs to the terpene pathway (125 clusters) (Figure 1). This biosynthesis pathway is well known to be present in many plant and fungi genomes, but recently it was proposed to be also widely distributed in bacterial genomes. One study revealed 262 distinct terpene synthases in the bacterial domain of life (Yamada et al., 2015). Consequently, it can represent a fertile source of new natural products, yet greatly underestimated. The second most abundant class of clusters belongs to the bacteriocin pathway (35 clusters) (Figure 1). bacteriocins belong to a group of ribosomal synthesized antimicrobial peptides which can kill or inhibit bacterial strains closely related or non-related to the bacteriocin producing bacteria (Yang et al., 2014). It has been suggested as a viable alternative to traditional antibiotics and can be used as narrow-spectrum antibiotics (Cotter et al., 2012). Only a few studies have been conducted to screen for bacteriocin genes by using a metagenomics approach, and solely for the host-associated microbiome (Zheng et al., 2015) or fermented food microbiome (Wieckowicz et al., 2011; Illegheems et al.,

2015; Escobar-Zepeda et al., 2016). Neither of these studies was conducted for natural environments or used the metagenomics 2.0 approach.

In this study, we focused on 2 families of large modular secondary metabolite genes, type I PKS and NRPS. With our approach, it was possible to find a total of 18 NRPS, 6 type I PKS, and 3 hybrid PKS/NRPS clusters. For NRPS clusters, it was possible to recover 43 C domains, most of them (58%) from the LCL class. An LCL domain catalyzes a peptide bond between two L-amino acids (Rausch et al., 2007). A previous study also found that the LCL class was the most abundant in another aquatic environment, dominated by gram-negative bacteria (Cuadrat et al., 2015). Many studies have shown that the LCL class in aquatic environments is limited to gram-negative bacteria (Woodhouse et al., 2013). Our results further support this as we also found the LCL class only in bins of gram-negative bacteria (Figure 3) with the only exception of unclassified Archaea (bin 233), which should be further investigated in order to confirm the phylogenetic classification. It was also possible to recover 9 KS domains from type I PKS clusters, 56% from the modular class and 22% from the hybrid PKS/NRPS class. Those classes are larger (with many copies of each domain) than the iterative ones, increasing the chances to be recovered by metagenomic approaches. Accordingly, the NRPS and PKS clusters were more in depth analyzed, including synteny, domain phylogeny, and partial metabolite protein structure predictions.



From the bins showing secondary metabolite genes, the most complete was from bin 34 (*Pseudomonas*). The 7 clusters on this genome vary from 8,675 base pairs (bp) to 52,516 bp in size, been only possible to be recovered due to the high completeness of the assembled genome (98.28%). The presence of a great diversity of clusters in *Pseudomonas* is expected, as many active secondary metabolites (encoded by NRPS) have been previously described in the *Pseudomonas* genus, ranging from antibiotics and antifungal to siderophores (Pan and Hu, 2015; Van Der Voort et al., 2015; Esmaeel et al., 2016).

From those 7 clusters in bin 34, the NRPS clusters 2 and 3 showed a high similarity with syringomycin (three domains) and microcystin (one domain) pathways. The first one is found, for example, in the *Pseudomonas syringae* (a plant pathogen) genome, as a virulence factor (syringomycin E) (Scholz-Schroeder et al., 2003), which also has antifungal activity against *Saccharomyces cerevisiae* (Stock et al., 2000). On the other hand, microcystin are a class of toxins produced by freshwater Cyanobacteria species (Dawson, 1998) and it can be produced in large quantities during massive bloom events (Bouhaddada et al., 2016). Due to the taxonomical classification of the bin and the higher number of domains similar to syringomycin, however, it is more likely that the product encoded by this cluster is functionally close to the latter pathway.

In bin 34-cluster 6, both C domains were classified as belonging to the pyochelin pathway. This peptide is a siderophore of *Pseudomonas aeruginosa* (Brandel et al., 2012). The presence of a TonB-dependent siderophore receptor in cluster 6 provides additional evidence about its functional classification. Additionally, two bacteriocin clusters and one Aryl polyene cluster were found in bin 34. Aryl polyenes are structurally similar to the well-known carotenoids with respect to their polyene systems and it was recently demonstrated that it can protect bacteria from reactive oxygen species, similarly

to what is known for carotenoids (Schöner et al., 2016). These results suggest that a wide range of metabolites is encoded in this *Pseudomonas* genome, providing it an “arsenal” of secondary products, increasing the likelihood of the *Pseudomonas* species to succeed in aquatic systems.

Bin 131 (unclassified bacteria) shows a PKS cluster and 3 domains. It was classified as belonging to the enediynes pathway. These compounds are toxic to DNA and are under investigation as anti-tumor agents, with several compounds under clinical trials (Jones and Fouad, 2002). All are encoded by type I iterative PKS (Ahlert et al., 2002) and it was possible to recover the minimal core (KS, AT and KR) as well as transporter genes from the environmental genome. The most similar PKS I present in the public databases stems from *M. aeruginosa*, but only with an identity of 64%, suggesting that it is encoding for a new compound, which has not previously been described.

In bin 193 (*Mycobacterium*—sister lineage *M. rhodesiae*), one of the 3 recovered clusters is a Type I PKS, similar to iterative PKS in the NAPDOS analysis (confirmed by the phylogenetic analysis). The most similar KS sequence belongs to *M. kansasii*, with 82% similarity. *M. rhodesiae* and *M. kansasii* are both non-tuberculous mycobacteria (NTM) that can be found in different environments, but both can also be opportunistic pathogens and cause a chronic pulmonary infection in immunosuppressed patients (Fedrizzi et al., 2017). The species *M. kansasii* comprises various subtypes and some are often recovered from tap water and occasionally from river or lake water (Bakuła et al., 2013; van der Wielen et al., 2013). There is still controversy about how the transmission from environment to human host occurs and also about the implications on public health. The presence of PKS in *Mycobacterium* genus was discovered more than a decade ago and most of the polyketides encoded by different species of this genus play a role in virulence and/or components of the extraordinarily complex mycobacterial cell

envelope (Quadri, 2014). Further studies must be done in order to investigate the potential of this bin to cause infections on humans, i.e., by screening virulence factors on the full genome.

Five bins of the phyla Cyanobacteria contained PKS and NRPS clusters. In bins 1 and 2 (*Anabaena*, now called “*Dolichospermum*”), it was possible to recover 2 NRPS and 1 hybrid NRPS-PKS, respectively. The genus *Anabaena* is known to encode several toxins, including the dangerous anatoxin-a, and to produce toxic blooms in lakes and reservoirs (Carmichael et al., 1975; Calteau et al., 2014; Brown et al., 2016; Li et al., 2016). However, the anatoxin-a is encoded by a type I PKS cluster (Méjean et al., 2014), unlike the NRPS and Hybrid clusters found in the *Anabaena* bins from this study.

On the other hand, the hepatotoxic heptapeptide of the class mycrocystin is present in many genera of Cyanobacteria, including *Anabaena*, and they are encoded by NRPS and also Hybrid NRPS-PKS clusters (Tillett et al., 2000; Rouhiainen et al., 2004; Viaggiu et al., 2004; Rastogi et al., 2015). The results of NAPDOS reveal one C domain from bin 1 classified in the pathway of Mycrocystin with e-value 6e-83.

In bins 6, 7 and 8 (*Planktothrix*), it was possible to find several type I PKS, NRPS and hybrid clusters. In bin 6 there are 3 NRPS, one type 1 PKS, and one hybrid cluster. The bin 7 shows 2 NRPS and bin 8 shows one type I PKS and 2 NRPS clusters. The genus *Planktothrix* can also be producer of anatoxin-a (Viaggiu et al., 2004) and the presence of type I PKS cluster on these bins can be alarming. However, the 3 KS domains from type I PKS from bin 6 reveal great similarity with the epothilone pathway. The NapDOS analysis and the KS domain from bin 8 suggest a high similarity to the neurotoxin jamaicamides pathway. A previous study showed in 2010 (Ballot et al., 2010a) the presence of an anatoxin-a-producing cyanobacterium in northeastern Germany Lake Stolpsee, rising concerns about the presence of these toxins in the waters of the northeastern German lakes.

The absence of anatoxin-a genes in the studied lake is in agreement with previous screening for a toxin screening in the lake (Dadheech et al., 2014).

In the bin 73 (*Acidobacteriales*) 1 PKS sequence was found. By the phylogenetic classification, its KS domain is clustered with *trans*-AT KS domains. The AT domains of *trans*-AT PKSs are not integrated into the assembly lines but expressed as free-standing polypeptides, unlike the more familiar *cis*-AT PKSs (Weissman, 2015). However, the NAPDOS result shows the AT domain of this bin in the same ORF with KS and KR domains, showing a syntenicity that suggests a *cis*-AT PKS. In addition, the classification by similarity from NAPDOS suggests a polyunsaturated fatty acid (PUFA) but only with 31% of identity.

To assess the life style of the bins (free-living or particle-associated), we calculated the relative abundance of the bins in every sample. A total of 158 bins with significant difference between the 3 groups were found (Table S7), however from the 15 bins on which this study focused (showing NRPS and/or type I PKS clusters), only 4 bins (26.6%) were significantly differently present in the life-styles. Bins 1 and 2 (*Anabaena* genus) are

more abundant in the PA group, especially on samples B7 and B9 (and also the replicates Old_b7 and Old_b9), accounting for 20–25% (bin 1) and 10–15% (bin 2) on these samples. The very high abundance of these bins on the samples can be explained by that, based on long term monitoring of the Lake Stechlin, we know that these samples were collected during the occurrence of a cyanobacterial bloom.

From the other bins containing PKS/NRPS clusters, we can see that Bins 6, 7, and 8 (*Planktothrix*), beside the lack of significant difference between FL and PA groups (p -value $> p$ 0.05), they are clearly more abundant in NSF. The possible explanation for this notion is that the NSF samples were collected during a mesocosm experiment, whereas the other samples were derived from the natural environment.

CONCLUSIONS

Using the Metagenomics 2.0 approach, we were able to recover full megasynthases sequences and their genomic context from environmental draft genomes. However, there are limitations, e.g., the genomic coverage of less abundant organisms and the possibility of chimeras. Recently, it has been demonstrated that with an increasing number of samples, it will be possible to recover individual species genomes with a high confidence (Kang et al., 2015). In the near future, with the advent of the 3rd generation sequencing, with longer reads, up to 100 kilobases, it will be possible to further improve the quality of the assemblies (Frank et al., 2016). These new approaches unlock the possibility of studying these newly recovered environmental pathways and their evolution in detail. Thus, allowing cloning and expressing these clusters will provide new natural products of great interest for the biotechnological and pharmaceutical industry. Moreover, studies have demonstrated the possibility to synthesize large functional DNA (Hutchison et al., 2016), and together with additional screening techniques, it will be possible to obtain such sequences and to synthesize the full cluster for heterologous expression, skipping the cloning and functional screening process, saving considerable time and money. In addition, the current work highlights the great potential for the discovery of new metabolically active compounds in freshwaters such as oligo-mesotrophic Lake Stechlin. Further, the study of complete or near complete genomes from uncultivated bacteria in the natural environment will enable us to better understand the multiple forms of interactions between species and how they compete for the limiting natural resources.

AUTHOR CONTRIBUTIONS

RC, DI, AD, and H-PG: Conceived and designed the experiments; RC, DI: Performed the experiments; RC, DI, and H-PG: Analyzed the data; DI and H-PG: Contributed reagents, materials, analysis tools; All authors wrote the manuscript and revised it for significant intellectual content.

FUNDING

This study was supported by the Science without Borders Program (Ciência Sem Fronteiras), CNPq. DI and H-PG were funded by German science foundation (DFG) projects Aquameth (GR1540/21-1) and Aggregates (GR1540/28-1).

ACKNOWLEDGMENTS

We thank Dr. Camila Mazzoni and all the team of Berlin Center for Genomics in Biodiversity Research (BeGenDiv) for allowing us to use the facilities and computational resources for the bioinformatics analyses. Elke Mach and the MIBI group are thanked for their technical support and fruitful discussions.

REFERENCES

- Ahlert, J., Shepard, E., Lomovskaya, N., Zazopoulos, E., Staffa, A., Bachmann, B. O., et al. (2002). The calicheamicin gene cluster and its iterative type I enediyne PKS. *Science* 297, 1173–1176. doi: 10.1126/science.1072105
- Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi: 10.1038/nbt.2579
- Amos, G. C., Borsetto, C., Laskaris, P., Krsek, M., Berry, A. E., Newsham, K. K., et al. (2015). Designing and implementing an assay for the detection of rare and divergent nrps and pks clones in European, Antarctic and Cuban soils. *PLoS ONE* 10:e0138327. doi: 10.1371/journal.pone.0138327
- Austin, M. B., and Noel, J. P. (2003). The chalcone synthase superfamily of type III polyketide synthases. *Nat. Prod. Rep.* 20, 79–110. doi: 10.1039/b100917f
- Bakula, Z., Safianowska, A., Nowacka-Mazurek, M., Bielecki, J., and Jagielski, T. (2013). Short communication: subtyping of *Mycobacterium kansasii* by PCR-restriction enzyme analysis of the hsp65Gene. *Biomed Res. Int.* 2013:178725. doi: 10.1155/2013/178725
- Ballot, A., Fastner, J., Lentz, M., and Wiedner, C. (2010a). First report of anatoxina-producing cyanobacterium *Aphanizomenon issatschenkoi* in northeastern Germany. *Toxicon* 56, 964–971. doi: 10.1016/j.toxicon.2010.06.021
- Ballot, A., Fastner, J., and Wiedner, C. (2010b). Paralytic shellfish poisoning toxin-producing cyanobacterium *Aphanizomenon gracile* in northeast Germany. *Appl. Environ. Microbiol.* 76, 1173–1180. doi: 10.1128/AEM.02285-09
- Bláhová, L., Babica, P., Adamovský, O., Kohoutek, J., Maršálek, B., and Bláha, L. (2008). Analyses of cyanobacterial toxins (microcystins, cylindrospermopsin) in the reservoirs of the Czech Republic and evaluation of health risks. *Environ. Chem. Lett.* 6, 223–227. doi: 10.1007/s10311-007-0126-x
- Bláhová, L., Babica, P., Maršálek, B., Smutná, M., Maršálek, B., and Bláha, L. (2007). Concentrations and seasonal trends of extracellular microcystins in freshwaters of the Czech Republic – results of the national monitoring program. *Clean Soil Air Water* 35, 348–354. doi: 10.1002/clen.200700010
- Bouhaddada, R., Néliu, S., Nasri, H., Delarue, G., and Bouaïcha, N. (2016). High diversity of microcystins in a microcystis bloom from an Algerian lake. *Environ. Pollut.* 216, 836–844. doi: 10.1016/j.envpol.2016.06.055
- Brandel, J., Humbert, N., Elhabiri, M., Schalk, I. J., Mislin, G. L. A., and Albrecht-Gary, A.-M. (2012). Pyochelin, a siderophore of *Pseudomonas aeruginosa*: physicochemical characterization of the iron(III), copper(II) and zinc(II) complexes. *Dalton Trans.* 41, 2820–2834. doi: 10.1039/c1dt11804h
- Brown, N. M., Mueller, R. S., Shepardson, J. W., Landry, Z. C., Morrè, J. T., Maier, C. S., et al. (2016). Structural and functional analysis of the finished genome of the recently isolated toxic *Anabaena* sp. WA102. *BMC Genomics* 17:457. doi: 10.1186/s12864-016-2738-7
- Calteau, A., Fewer, D. P., Latif, A., Coursin, T., Laurent, T., Jokela, J., et al. (2014). Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* 15:977. doi: 10.1186/1471-2164-15-977

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2018.00251/full#supplementary-material>

Table S1 | Samples used in the study.

Table S2 | Taxonomical classification of bins, GC content and genome size.

Table S3 | Secondary metabolite clusters found in each bin.

Table S4 | KS domains blastP results against Refseq. Only top 3 hits are shown.

Table S5 | C domains blastP results against Refseq. Only top 3 hits are shown.

Table S6 | Relative abundance of each bin over the samples.

Table S7 | Stamp results. Only the bins where relative abundance was statistically significant different ($p < 0.05$) between the 3 groups (FL, PA, and NSF) are shown.

- Cane, D. E., Walsh, C. T., and Khosla, C. (1998). Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* 282, 63–68. doi: 10.1126/science.282.5386.63
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303
- Carmichael, W. W., Biggs, D. F., and Gorham, P. R. (1975). Toxicology and pharmacological action of *Anabaena flos-aquae* toxin. *Science* 187, 542–544.
- Cotter, P. D., Ross, R. P., and Hill, C. (2012). Bacteriocins — a viable alternative to antibiotics? *Nat. Rev. Microbiol.* 11, 95–105. doi: 10.1038/nrmicro2937
- Cuadrat, R. R., Cury, J. C., and Dávila, A. M. (2015). Metagenomic analysis of upwelling-affected Brazilian coastal seawater reveals sequence domains of type I PKS and modular NRPS. *Int. J. Mol. Sci.* 16, 28285–28295. doi: 10.3390/ijms161226101
- Dadheech, P. K., Selmezy, G. B., Vasas, G., Padisák, J., Arp, W., Tapolczai, K., et al. (2014). Presence of potential toxin-producing cyanobacteria in an Oligo-Mesotrophic Lake in Baltic Lake district, Germany: an ecological, genetic and toxicological survey. *Toxins* 6, 2912–2931. doi: 10.3390/toxins6102912
- Dawson, R. M. (1998). The toxicology of microcystins. *Toxicon* 36, 953–962. doi: 10.1016/S0041-0101(97)00102-5
- Di Lorenzo, M., Stork, M., Naka, H., Tolmasky, M. E., and Crosa, J. H. (2008). Tandem heterocyclization domains in a nonribosomal peptide synthetase essential for siderophore biosynthesis in *Vibrio anguillarum*. *Biomaterials* 21, 635–648. doi: 10.1007/s10534-008-9149-4
- Escobar-Zepeda, A., Sanchez-Flores, A., and Quirasco Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiol.* 57, 116–127. doi: 10.1016/j.fm.2016.02.004
- Esmaeel, Q., Pupin, M., Kieu, N. P., Chataign, G., Béchet, M., Deravel, J., et al. (2016). *Burkholderia* genome mining for nonribosomal peptide synthetases reveals a great potential for novel siderophores and lipopeptides synthesis. *Microbiologyopen* 5, 512–526. doi: 10.1002/mbo3.347
- Fedrizzi, T., Meehan, C. J., Grottole, A., Giacobazzi, E., Serpini, G. F., Tagliacuzzi, S., et al. (2017). Genomic characterization of nontuberculous mycobacteria. *Nature* 7:45258. doi: 10.1038/srep45258
- Fisch, K. M. (2013). Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. *RSC Adv.* 3, 18228–18247. doi: 10.1039/c3ra42661k
- Foerster, K. U., Doerks, T., Creevey, C. J., Doerks, A., and Bork, P. (2008). A computational screen for type I polyketide synthases in metagenomics shotgun data. *PLoS ONE* 3:e3515. doi: 10.1371/journal.pone.0003515
- Frank, J. A., Pan, Y., Tooming-Klunderud, A., Eijsink, V. G. H., McHardy, A. C., Nederbragt, A. J., et al. (2016). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci. Rep.* 6:25373. doi: 10.1038/srep25373
- Gilbert, J. A., Meyer, F., and Bailey, M. J. (2011). The future of microbial metagenomics (or is ignorance bliss?). *ISME J. Int. Soc. Microb. Ecol.* 5, 777–779. doi: 10.1038/ismej.2010.178

- Gokhale, R. S., Sankaranarayanan, R., and Mohanty, D. (2007). Versatility of polyketide synthases in generating metabolic diversity. *Curr. Opin. Struct. Biol.* 17, 736–743. doi: 10.1016/j.sbi.2007.08.021
- Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107
- Hutchison, C. A. III., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., et al. (2016). Design and synthesis of a minimal bacterial genome. *Science* 351:aad6253. doi: 10.1126/science.aad6253
- Illegheems, K., Weckx, S., and De Vuyst, L. (2015). Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample. *Food Microbiol.* 50, 54–63. doi: 10.1016/j.fm.2015.03.005
- Ionescu, D., Siebert, C., Polerecky, L., Munwes, Y. Y., Lott, C., Häusler, S., et al. (2012). Microbial and chemical characterization of underwater fresh water springs in the Dead Sea. *PLoS ONE* 7:e38319. doi: 10.1371/journal.pone.0038319
- Jones, G. B., and Fouad, F. S. (2002). Designed enediynes antitumor agents. *Curr. Pharm. Des.* 8, 2415–2440. doi: 10.2174/1381612023392810
- Jost Casper, S. (ed.). (1985). *Lake Stechlin: A Temperate Oligotrophic Lake (Monographiae Biologicae), 1st Edn.*
- Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. doi: 10.7717/peerj.1165
- Kirillin, G., Shatwell, T., and Kasprzak, P. (2013). Consequences of thermal pollution from a nuclear plant on lake temperature and mixing regime. *J. Hydrol.* 496, 47–56. doi: 10.1016/j.jhydrol.2013.05.023
- Koglin, A., and Walsh, C. T. (2009). Structural insights into nonribosomal peptide enzymatic assembly lines. *Nat. Prod. Rep.* 26, 987–1000. doi: 10.1039/b904543k
- Komaki, H., Ichikawa, N., Hosoyama, A., Takahashi-Nakaguchi, A., Matsuzawa, T., Suzuki, K., et al. (2014). Genome based analysis of type-I polyketide synthase and nonribosomal peptide synthetase gene clusters in seven strains of five representative *Nocardia* species. *BMC Genomics* 15:323. doi: 10.1186/1471-2164-15-323
- Lal, R., Kumari, R., Kaur, H., Khanna, R., Dhingra, N., and Tuteja, D. (2000). Regulation and manipulation of the gene clusters encoding type-I PKSs. *Trends Biotechnol.* 18, 264–274. doi: 10.1016/S0167-7799(00)01443-8
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033
- Li, X., Dreher, T. W., and Li, R. (2016). An overview of diversity, occurrence, genetics and toxin production of bloom-forming *Dolichospermum* (*Anabaena*) species. *Harmful Algae* 54, 54–68. doi: 10.1016/j.hal.2015.10.015
- Masschelein, J., Mattheus, W., Gao, L.-J., Moons, P., Van Houdt, R., Uytterhoeven, B., et al. (2013). A PKS/NRPS/FAAS hybrid gene cluster from *Serratia plymuthica* RVH1 encoding the biosynthesis of three broad spectrum, zeamine-related antibiotics. *PLoS ONE* 8:e54143. doi: 10.1371/journal.pone.0054143
- McMahon, K. (2015). 'Metagenomics 2.0.' *Environ. Microbiol. Rep.* 7, 38–39. doi: 10.1111/1758-2229
- Méjean, A., Paci, G., Gautier, V., and Ploux, O. (2014). Biosynthesis of anatoxin-a and analogues (anatoxins) in cyanobacteria. *Toxicon* 91, 15–22. doi: 10.1016/j.toxicon.2014.07.016
- Meriluoto, J., Spoof, L., and Codd, G. A. (eds.). (2017). *Handbook of Cyanobacterial Monitoring and Cyanotoxin Analysis*. Chichester: Wiley.
- Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386
- Micallef, M. L., D'Agostino, P. M., Sharma, D., Viswanathan, R., and Moffitt, M. C. (2015). Genome mining for natural product biosynthetic gene clusters in the Subsection V cyanobacteria. *BMC Genomics* 16:669. doi: 10.1186/s12864-015-1855-z
- Minowa, Y., Araki, M., and Kanehisa, M. (2007). Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* 368, 1500–1517. doi: 10.1016/j.jmb.2007.02.099
- Mizuno, C. M., Kimes, N. E., López-Pérez, M., Ausó, E., Rodríguez-Valera, F., and Ghai, R. (2013). A hybrid NRPS-PKS gene cluster related to the bleomycin family of antitumor antibiotics in *Alteromonas macleodii* strains. *PLoS ONE* 8:e76021. doi: 10.1371/journal.pone.0076021
- National Research Council US. (2007). *Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, DC: National Academies Press (US); Data Management and Bioinformatics Challenges of Metagenomics.
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189
- Paerl, H. W., and Huisman, J. (2008). Blooms like it hot. *Science* 320, 57–58. doi: 10.1126/science.1155398
- Pan, H.-Q., and Hu, J.-C. (2015). Draft genome sequence of the novel strain *Pseudomonas* sp. 10B238 with potential ability to produce antibiotics from deep-sea sediment. *Mar. Genomics* 23, 55–57. doi: 10.1016/j.margen.2015.05.003
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114
- Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123–3124. doi: 10.1093/bioinformatics/btu494
- Quadri, L. E. (2014). Biosynthesis of mycobacterial lipids by polyketide synthases and beyond. *Crit. Rev. Biochem. Mol. Biol.* 49, 179–211. doi: 10.3109/10409238.2014.896859
- Rantala, A., Rizzi, E., Castiglioni, B., de Bellis, G., and Sivonen, K. (2008). Identification of hepatotoxin producing cyanobacteria by DNA-chip. *Environ. Microbiol.* 10, 653–664. doi: 10.1111/j.1462-2920.2007.01488.x
- Rastogi, R. P., Madamwar, D., and Incharoensakdi, A. (2015). Bloom dynamics of cyanobacteria and their toxins: environmental health impacts and mitigation strategies. *Front. Microbiol.* 6:1254. doi: 10.3389/fmicb.2015.01254
- Rausch, C., Hoof, I., Weber, T., Wohlleben, W., and Huson, D. H. (2007). Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* 7:78. doi: 10.1186/1471-2148-7-78
- Rodríguez-Valera, F. (2004). Environmental genomics, the big picture? *FEMS Microbiol. Lett.* 231, 153–158. doi: 10.1016/S0378-1097(04)00006-0
- Rouhiainen, L., Vakkilainen, T., Siemer, B. L., Buikema, W., Haselkorn, R., and Sivonen, K. (2004). Genes coding for hepatotoxic heptapeptides (Microcystins) in the cyanobacterium *Anabaena* strain 90. *Appl. Environ. Microbiol.* 70, 686–692. doi: 10.1128/AEM.70.2.686-692.2004
- Salmaso, N., and Padisák, J. (2007). Morpho-functional groups and phytoplankton development in two deep lakes (Lake Garda, Italy and Lake Stechlin, Germany). *Hydrobiologia* 578, 97–112. doi: 10.1007/s10750-006-0437-0
- Sangwan, N., Xia, F., and Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4:8. doi: 10.1186/s40168-016-0154-5
- Scholz-Schroeder, B. K., Soule, J. D., and Gross, D. C. (2003). The *sypA*, *sypB*, and *sypC* synthetase genes encode twenty-two modules involved in the nonribosomal peptide synthesis of syringopeptin by *Pseudomonas syringae* pv. *syringae* B301D. *Mol. Plant Microbe Inter.* 16, 271–280. doi: 10.1094/MPMI.2003.16.4.271
- Schöner, T. A., Gassel, S., Osawa, A., Tobias, N. J., Okuno, Y., Sakakibara, Y., et al. (2016). Aryl polyenes, a highly abundant class of bacterial natural products, are functionally related to antioxidative carotenoids. *Chembiochem* 17, 247–253. doi: 10.1002/cbic.201500474
- Selvin, J., Sathiyarayanan, G., Lipton, A. N., Al-Dhabi, N. A., Valan Arasu, M., and Kiran, G. S. (2016). Ketide Synthase (KS) domain prediction and analysis of iterative type II PKS gene in marine sponge-associated actinobacteria producing biosurfactants and antimicrobial agents. *Front. Microbiol.* 7:63. doi: 10.3389/fmicb.2016.00063
- Silva-Stenico, M. E., Silva, C. S. P., Lorenzi, A. S., Shishido, T. K., Etcheagaray, A., Lira, S. P., et al. (2011). Non-ribosomal peptides produced by Brazilian cyanobacterial isolates with antimicrobial activity. *Microbiol. Res.* 166, 161–175. doi: 10.1016/j.micres.2010.04.002
- Stock, S. D., Hama, H., Radding, J. A., Young, D. A., and Takemoto, J. Y. (2000). Syringomycin E inhibition of *Saccharomyces cerevisiae*: requirement for biosynthesis of sphingolipids with very-long-chain fatty acids and

- mannose-and phosphoinositol-containing head groups. *Antimicrob. Agents Chemother.* 44, 1174–1180. doi: 10.1128/AAC.44.5.1174-1180.2000
- Strous, M., Kraft, B., Bisdorf, R., and Tegetmeyer, H. E. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front. Microbiol.* 3:410. doi: 10.3389/fmicb.2012.00410
- Sun, W., Peng, C., Zhao, Y., and Li, Z. (2012). Functional gene-guided discovery of type II polyketides from culturable actinomycetes associated with soft coral *Scleronephthya* sp. *PLoS ONE* 7:e42847. doi: 10.1371/journal.pone.0042847
- Teeling, H., and Glöckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Brief. Bioinformatics* 13, 728–742. doi: 10.1093/bib/bbs039
- Tillett, D., Dittmann, E., Erhard, M., von Döhren, H., Börner, T., and Neilan, B. A. (2000). Structural organization of microcystin biosynthesis in *Microcystis aeruginosa* PCC7806: an integrated peptide-polyketide synthetase system. *Chem. Biol.* 7, 753–764. doi: 10.1016/S1074-5521(00)00021-1
- Van Der Voort, M., Meijer, H. J. G., Schmidt, Y., Watrous, J., Dekkers, E., Mendes, R., et al. (2015). Genome mining and metabolic profiling of the rhizosphere bacterium *Pseudomonas* sp. SH-C52 for antimicrobial compounds. *Front. Microbiol.* 6:693. doi: 10.3389/fmicb.2015.00693
- van der Wielen, P. W., Heijnen, L., and van der Kooij, D. (2013). Pyrosequence analysis of the hsp65 genes of nontuberculous mycobacterium communities in unchlorinated drinking water in the Netherlands. *Appl. Environ. Microbiol.* 79, 6160–6166. doi: 10.1128/AEM.01591-13
- Vanwongerghem, I., Jensen, P. D., Rabaey, K., and Tyson, G. W. (2016). Genome-centric resolution of microbial diversity, metabolism and interactions in anaerobic digestion: genome-centric resolution through deep metagenomics. *Environ. Microbiol.* 18, 3144–3158. doi: 10.1111/1462-2920.13382
- Viaggiu, E., Melchiorre, S., Volpi, F., Di Corcia, A., Mancini, R., Garibaldi, L., et al. (2004). Anatoxin-a toxin in the cyanobacterium *Planktothrix rubescens* from a fishing pond in northern Italy. *Environ. Toxicol.* 19, 191–197. doi: 10.1002/tox.20011
- Vincent, A. T., Derome, N., Boyle, B., Culley, A. I., and Charette, S. J. (2016). Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *J. Microbiol. Methods.* 138, 60–71. doi: 10.1016/j.mimet.2016.02.016
- Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Brucoleri, R., et al. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 43, W237–W243. doi: 10.1093/nar/gkv437
- Weissman, K. J. (2015). The structural biology of biosynthetic megaenzymes. *Nat. Chem. Biol.* 11, 660–670. doi: 10.1038/nchembio.1883
- Wieckowicz, M., Schmidt, M., Sip, A., and Grajek, W. (2011). Development of a PCR-based assay for rapid detection of class IIa bacteriocin genes: detection of class IIa bacteriocins. *Lett. Appl. Microbiol.* 52, 281–289. doi: 10.1111/j.1472-765X.2010.02999.x
- Woodhouse, J. N., Fan, L., Brown, M. V., Thomas, T., and Neilan, B. A. (2013). Deep sequencing of non-ribosomal peptide synthetases and polyketide synthases from the microbiomes of Australian marine sponges. *ISME J.* 7, 1842–1851. doi: 10.1038/ismej.2013.65
- Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2:26. doi: 10.1186/2049-2618-2-26
- Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-ya, K., Omura, S., Cane, D. E., et al. (2015). Terpene synthases are widely distributed in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 112, 857–862. doi: 10.1073/pnas.1422108112
- Yang, S.-C., Lin, C.-H., Sung, C. T., and Fang, J.-Y. (2014). Antibacterial activities of bacteriocins: application in foods and pharmaceuticals. *Front. Microbiol.* 5:241. doi: 10.3389/fmicb.2014.00241
- Zheng, J., Gänzle, M. G., Lin, X. B., Ruan, L., and Sun, M. (2015). Diversity and dynamics of bacteriocins from human microbiome: bacteriocins of human microbiome. *Environ. Microbiol.* 17, 2133–2143. doi: 10.1111/1462-2920.12662
- Ziemert, N., Podell, S., Penn, K., Badger, J. H., Allen, E., and Jensen, P. R. (2012). The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE* 7:e34064. doi: 10.1371/journal.pone.0034064
- Zothanpuia, Passari, A. K., Gupta, V. K., and Singh, B. P. (2016). Detection of antibiotic-resistant bacteria endowed with antimicrobial activity from a freshwater lake and their phylogenetic affiliation. *PeerJ* 4:e2103. doi: 10.7717/peerj.2103

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Cuadrat, Ionescu, Dávila and Grossart. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.